Utterance fluency development across CEFR-based proficiency assessments from A2 to B2

Elina Vasu

Tampere University and University of Jyväskylä, Finland

This article reports a study that investigated typical utterance fluency features' development across three oral proficiency levels (A2—B2) as measured by a local application of the Common European Framework of Reference for languages. The speakers were 60 teenaged learners of English from Finland. Approximately 20 seconds of their semi-spontaneous speech was analysed for eleven speech features related to speed, breakdown, and repair fluency. Results reveal that the speakers' tempo, number of words produced and length of uninterrupted speech between pauses increased along with proficiency. Also, the number of silent pauses in unconventional (mid-clause) positions and length of silent pauses decreased along with increased proficiency. As an implication, second-language learners could benefit from fluency practice such as focus on speed of delivery and pausing. Moreover, these aspects could be considered more consistently in language proficiency scales.

Keywords: second-language (L2) assessment, second-language (L2) oral proficiency, utterance fluency, Common European Framework of Reference for languages (CEFR)

Introduction

Speaking, as well as the flip side of the coin, listening, are utterly important skills to second and foreign language (L2) learners. Even though reading and writing skills are also needed, speaking is more crucial for daily activities such as buying items, building relationships, and finding employment. Communication via letters and emails has decreased; instead, video chat and video conferencing have increased both on free time

---

Email address for correspondence: elina.vasu@tuni.fi

and in work contexts. The importance of spoken communication is also acknowledged in modern L2 teaching, where emphasis is put on oral activities, offering students opportunities to practise speaking situations in class. In addition, students' speaking skills are tested with oral exams, and many high-stakes exams, too, include assessment of speaking skills. As for the goals, native-likeness is no longer something to strive for, but a more common aim is to produce the target language intelligibly and fluently.

Fluency is a complex concept to define. In its broad sense, fluency can be understood as a synonym for overall language proficiency (Ghanem & Kang, 2018), but in the present study, a narrow understanding of fluency is adopted, with focus on utterance fluency (Segalowitz, 2010). Utterance fluency refers to the measurable aspects of speech fluency, which can be classified into speed (e.g., speech and articulation rate), breakdown (e.g., pauses), and repair (e.g., repetitions) fluency (Skehan, 2009). Research demonstrates that fluent speech is associated with faster delivery, fewer and shorter pauses, and pauses placed at clause boundaries instead of in the middle of clauses (e.g., Kormos & Dénes, 2004; Pinget et al., 2014). In addition, it has been suggested that self-corrections and filled pauses (e.g., um, uh, er) may also link with fluent speech; that is, self-correction is more fluent than making errors, and so are filled pauses compared to (long) silent pauses (see e.g., Peltonen, 2018). Typical utterance fluency features have also been studied in connection to L2 comprehensibility, i.e., "the ease or difficulty a listener experiences in understanding an utterance" (Derwing & Munro, 2015, p. 5). As can be presumed, many speech features that are associated with fluency have been found associated with comprehensibility as well. These include particularly speech rate (e.g., Munro & Derwing, 1998, 2001; Saito et al., 2016, 2017; Trofimovich & Baker, 2006; Volín & Skarnitzl, 2010), pausing (e.g., Kang, 2010; Trofimovich & Baker, 2006), and length of runs between pauses (e.g., Kang, 2010; Trofimovich & Isaacs, 2012).

As native-likeness is no longer an overarching learning goal for L2 speech (see Levis, 2005), utterance fluency, comprehensibility, and intelligibility ("the degree of match between a speaker's intended message and the listener's comprehension", Derwing & Munro, 2015, p. 5) are taken into account in L2 testing: several proficiency scales address them in their assessment criteria, alongside with various other aspects of language (see the Utterance Fluency in Language Proficiency Scales subsection below).

Yet the role of utterance fluency in L2 proficiency assessments calls for further research. Namely, previous studies have yielded mixed results on the link between utterance fluency and proficiency assessments; for example, filled pauses such as *um, uh, er* have been linked with proficiency assessments in some studies (e.g., Kang & Yan, 2018) but not in others (e.g., Ginther et al., 2010; more examples are discussed in the subsection below: Previous studies on how utterance fluency contributes to proficiency ratings). In addition, commonly used L2 testing instruments have varied conceptualisations and operationalisations of fluency (Huhta et al., 2020), which is a further complication when striving for an understanding of the role of utterance fluency in L2 proficiency assessments.

Further investigation into the role of utterance fluency in L2 proficiency assessments is needed to re-evaluate the ways utterance fluency is described in proficiency scales. Harding (2017) critiqued the Common European Framework of Reference for languages (CEFR; Council of Europe, 2001) regarding the phonological control scale which was soon after revised and published in a companion volume (Council of Europe, 2020). This example shows how important research is for the ongoing development of L2 testing. Research is particularly important for the development of automated language assessment. Traditionally, automatic speech recognition (ASR) technologies used in automated language assessment were trained to follow predefined rules. To do this, the rules need to be first established. More recently, deep neural network (DNN) algorithms have been developed for the training. These algorithms can learn patterns from the data, but they are yet not perfect (Evanini & Zechner, 2020), and according to some, even the "state-of-the-art automated scoring algorithms do not and cannot think like humans" (Zhang et al., 2020, p. 23). To conclude, even the newest technological advancements have not set aside the need to further investigate the role of different speech features in L2 proficiency assessments.

The current study seeks to present new evidence on the role of utterance fluency in L2 proficiency assessments. The question is addressed with a research design that utilises a novel speaker group (L1 Finnish and Finland-Swedish teenagers learning English) and L2 proficiency assessments according to a local application of the CEFR (Finnish National Agency for Education, 2004). To be clear about the concepts of the present study, *proficiency* refers to these (and other) L2 proficiency assessments throughout the article, and *(utterance) fluency* and the measurable components of it such as *speed* are not considered synonyms of proficiency.

Second-language speech features often derive from the speakers' L1 (see e.g., Gass & Selinker, 1992). Hence, it is important to include less-researched L1 groups in this line of research and to seek verification for results obtained with other L1 speaker groups. In addition, the number of studies conducted on the CEFR and especially its local applications is still modest, as many previous studies have focussed on other proficiency scales such as the International English Language Testing System (IELTS, e.g., Kang et al., 2021), Test of English as a Foreign Language (TOEFL; e.g., Zechner et al., 2009), and the Oral English Proficiency Test (OEPT; e.g., Ginther et al., 2010). In sum, the present study will offer an interesting context to investigate the topic, using a CEFR-based local rating scale and a less-researched speaker group, both in terms of age and L1.

Literature review

Utterance fluency in language proficiency scales

Overall, the concept of fluency is present in many language proficiency scales. Hence, it is considered an important component in the assessment of L2 proficiency. Huhta and colleagues' (2020) systematic review of proficiency scales reveals that the fluency criterion is not always named 'fluency', but it can also be called 'discourse management', 'content and response', or 'clarity and naturalness of speech', for example. Regardless of the name, the criterion often considers typical utterance (dis)fluency features such as pauses, speed, length of expression, and hesitation (Huhta et al., 2020). However, the features may not be considered very systematically across proficiency levels. Pennington and Rogerson-Revell (2019) have noted that the descriptors of the American Council on the Teaching of Foreign Languages Oral Proficiency Interview (ACTFL OPI) have great variation within and across levels in how fluency, intelligibility, accuracy and accent are described, which is a similar critique to that presented in Harding (2017) towards the original CEFR from 2001. I will next describe fluency in the CEFR, as the present study is focussed on CEFR-based proficiency assessments.

Fluency is considered very vaguely in the original global CEFR scale. Proficient language users (C1–C2) can express themselves “spontaneously, very fluently and precisely” (C2) or “fluently and spontaneously without much obvious searching for expressions” (C1) (Council of Europe, 2001, p. 24). As for independent language users (B1–B2), level B2 speakers are described as being able to “interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party” (ibid.), whereas fluency as such is not mentioned in the descriptors for proficiency level B1 and lower. However, a more detailed assessment grid presenting the qualitative aspects of spoken language use offers more details about fluency at different proficiency levels. As for the proficient users, the grid adds expressing ideas “at length with natural colloquial flow” (C2) and “almost effortlessly” (C1) (ibid., p. 28). In addition, the word *smooth* is used to describe the fluency of proficient speakers. Independent language users are described in terms of tempo and hesitation: B2 speakers can produce language “with a fairly even tempo” but can be “hesitant”; however, their speech contains “few noticeably long pauses” (ibid., p. 28). In B1 speakers, “pausing for grammatical and lexical planning and repair is evident” (ibid., p. 29). As for basic users (A1–A2), the descriptors focus on pausing, false starts and reformulation, which are evident in A2 speakers (ibid., p. 29).

The CEFR companion volume (Council of Europe, 2020) brought no essential changes to the consideration of fluency. The scale of overall oral interaction describes proficient speakers in terms of fluency, spontaneity, and effortlessness, with the addition that “there is little obvious searching for expressions or avoidance strategies; only a conceptually difficult subject can hinder a natural, smooth flow of language” (C1) (ibid., p. 72). Fluency and spontaneity characterise independent language users (B2) as well. Unlike in the original CEFR, more detailed descriptions of fluency are presented in a designated fluency scale. Compared to the original CEFR, pausing forms a clear continuum from “much pausing” (A1) → pauses being “very evident” (A2) → “some problems with formulation resulting in pauses” (B1) → “few noticeably long pauses” (B2) → speakers pausing “only to reflect on precisely the right means to express their thoughts or to find an appropriate example or explanation” (C2) (ibid., p. 142).

Despite the somewhat vague and unsystematic descriptions of fluency in L2 proficiency scales such as the CEFR, fluency is a key aspect in L2 assessment, as characterised by Huhta et al. (2020). Even if named, defined, and operationalised differently in different proficiency scales, fluency is often considered in assessments. Consequently, utterance fluency features can distinguish proficiency levels, even though there may be differences in more features between high and low achievers than between adjacent proficiency levels (e.g., Kang & Yan, 2018). I will next explore utterance fluency features which have been frequently considered in research that aims to map which features contribute to proficiency ratings.

Previous studies on how utterance fluency contributes to proficiency ratings

Several scholars have been interested in the possible link between L2 proficiency assessments and utterance fluency. Some of them have conducted fine-grained analyses on how individual utterance fluency features correlate with proficiency assessments, while others have grouped features. Reviewing the literature, it is possible to identify features that are closely associated with proficiency assessments but also features that have yielded mixed results. Next, I will discuss previous studies from the viewpoint of the link; whether previous research has verified the association, or whether the findings have been mixed.

Starting from the verified links, scholars seem to be united on speech and articulation rate, and token frequency (i.e., total number of words produced) being associated with L2 proficiency assessments. Regarding token frequency, a clear link with proficiency assessments was detected by Kang and Yan's (2018) study on test takers of the CEFR-based Cambridge English Language Assessments (CELA) and Brown's (2006) study on IELTS examinees. Brown's study also reports close to significant ($p = .054$) differences in speech rate across proficiency levels; however, Brown excluded repairs and restarts from the speaking time. Other studies report clear links between speech rate and/or articulation rate and proficiency level (e.g., Ginther et al., 2010; Handley & Wang, 2024; Iwashita et al., 2008; Kang & Yan, 2018; Tavakoli et al., 2020). Also, Kang and Johnson's (2018) computer model found that proficiency assessments were affected by speech rate measures. Hence, there is already strong evidence of token frequency and tempo (measured as speech or articulation rate) being associated with

L2 proficiency assessments. Yet, these aspects are addressed in the present study for further verification.

Interestingly, some utterance fluency features have yielded mixed results in previous studies exploring the role of fluency in L2 proficiency assessments. I will first discuss mean length of run (MLR) and measurements of pausing. In Kang and Yan (2018), CELA test takers with better proficiency assessments spoke in longer runs, the link between MLR and proficiency assessment being statistically significant. The same result was obtained in Ginther and colleagues' (2010) study on OEPT examinees and Tavakoli et al. (2020), who studied speakers taking the British Council's Aptis Speaking Test. Then again, Iwashita et al. (2008) did not find a significant difference across proficiency levels for mean length of run. As for pausing measurements, Kang and Yan (2018) discovered that higher proficiency assessments were linked with fewer and shorter pauses in speech – both silent and filled (*um, uh, er*). Tavakoli et al. (2020) only found a link with filled pauses, whereas Ginther et al. (2010) and Iwashita et al. (2008) discovered the opposite: in their studies, a link was detected between proficiency level and silent pauses only. They also found a significant link between proficiency assessments and silent pause length and total pause time, respectively. Then again, Handley and Wang (2024) and Tavakoli et al. (2020) found mid-clause pause frequency linked with proficiency level. As for pause duration, the studies are conflicting: Tavakoli et al. (2020) detected a link with proficiency level, but Handley and Wang (2024) did not. Curiously, Brown (2006) did not detect statistically significant differences in any pausing measurement across proficiency levels. Similarly, Kang and Johnson's (2018) computer model did not find a direct link between proficiency level and silent pauses, and the correlation between proficiency level and filled pauses was weak. They conclude that filled pauses may not be a strong indicator of proficiency, which is in line with the findings of Iwashita et al. (2008) and Ginther et al. (2010).

Previous studies on hesitation phenomena and repairs have also yielded mixed results. In a study by Iwashita et al. (2008), repair was used as an umbrella term for repetitions, replacements, reformulations, and false starts. These measurements were merged and studied for a connection to TOEFL iBT scores, resulting in nonsignificant differences across proficiency levels. In Brown (2006), the analysis focused on and merged restarts and repetitions, not finding these self-corrections significantly linked with proficiency assessments. However, Handley and Wang's (2024) recent study on IELTS test takers suggests that repetition frequency may predict oral proficiency assessments, and Tavakoli et al. (2020) discovered that the occurrence of false starts and reformulations were statistically different across proficiency levels. As the number of studies regarding hesitation and repair is still low, no definite conclusions can be made. Consequently, hesitation and repair are investigated in the present study to widen the body of research on this topic.

To summarise, previous findings on the contribution of utterance fluency features to oral proficiency assessments are mixed. Clearest results have been reported regarding speed of delivery, i.e., speech rate and articulation rate, and it is safe to say that L2 proficiency assessment "rewards" speakers for development in speed fluency – even when assessed by the CEFR, which only mentions 'even tempo' in the descriptors. Token frequency is another feature that seems to have a clear link to proficiency assessments. When other utterance fluency features are concerned, the number of studies is still low and both positive and negative findings have been presented regarding several features. Hence, definitive conclusions cannot be made but more research is needed before drawing conclusions.

L1 Finnish and Finland-Swedish learners' utterance fluency of L2 English

The geographical context of the present study is Finland, which sets an interesting scene for this investigation. English has no official status in Finland, but the world-known school system offers extensive opportunities to study it as a foreign language, starting from Year 1, when students are six to seven years of age. English is not a compulsory subject as such, but it is obligatory to study at least one foreign language. In many primary schools, however, English is the only foreign language offered; hence, most students begin their English studies at the age of six or seven. In schools where children can start their language studies in Year 1 with a language other than English (e.g., Spanish, French, or German), English is studied as an elective from Year 4, when students are typically nine to ten years old. At the age of 15–16, students finish their nine-year basic education and should reach the English proficiency of an independent language user, i.e., B1.1 on the Finnish application of the CEFR scale (Finnish National Agency for Education, 2014). Due to immigration, students have varied linguistic and cultural backgrounds, but this study is limited to Year 9 students with the official languages of Finland as their L1: Finnish and Swedish. Even though the Swedish-speaking population makes out only 5 per cent of Finland's population and are often functionally bilingual in Swedish and Finnish, the country is officially bilingual.

Investigating teenaged speakers was seen desirable, because many previous fluency studies have focussed on adult speakers (e.g., De Jong et al., 2015; Pinget, et al., 2014), often overlooking students with low proficiency (e.g., Kang & Johnson, 2018; Kang & Yan, 2018). Especially in English as a foreign language (EFL) settings, language learners receiving formal education (and whose language skills are also assessed) are mainly adolescents. Focussing on L2 speakers of English from Finland aims to widen the range of L1s in research on the role of fluency in L2 proficiency assessments. It is important to study new speaker groups because speech features may surface differently based on the speakers' L1. For example, certain features can be more prevalent in one L1 than another, causing differences in the overall role of the feature on assessments across proficiency levels (cf. Tergujeff-Vasu, forthcoming).

Previous research on Finns' L2 English utterance fluency is scarce, but some knowledge has been obtained. Listener perception studies suggest that L1 Finnish youth can be perceived to speak English too slowly, with frequent hesitation and pauses at inconvenient places (Paananen-Porkka, 2007; Pihko, 1997). Slow speech rate and unusual pausing have been verified by Peltonen (2018), who studied L2 English utterance fluency in L1 Finnish Year 9 students and upper secondary school Year 1 students. Similarly, Tergujeff-Vasu (forthcoming) discovered that L1 Finnish and L1 Finland-Swedish students spoke English slower than the optimal rates suggested by Munro and Derwing (2001). Overall, it seems that Finns' English may be characterised by frequent, lengthy pauses, short runs between those pauses, and slow speech (Lehtonen, 1979; Peltonen, 2018; Tergujeff-Vasu, forthcoming), which all have negative effects on L2 fluency. Hence, it will be interesting to see how these issues are reflected in the students' L2 proficiency assessments.

The present study

Aim and research question

Since previous research results are mainly mixed, the current study seeks to present new evidence on the link between utterance fluency and L2 proficiency assessments. The following research question was set for the study:

> *How are typical utterance fluency features linked with L2 proficiency assessments based on the Finnish application of the CEFR?*

To address the research question, speech samples elicited from learners with A2, B1 and B2 proficiency in L2 English were analysed for eleven typical utterance fluency features. The measurements were contrasted across the proficiency levels, seeking links between the features and the assessments. Next, the research design is discussed in more detail, including data, participants, and analyses.

Data and participants

To acquire speech data, the author collaborated with the Finnish National Agency for Education and the Finnish Education Evaluation Centre. They provided access to their language assessment data, which was collected to evaluate Year 9 students' L2 English skills on a national level (Härmälä et al., 2014). As part of this national assessment, students' spoken proficiency was tested with a series of four tasks, which were video recorded at the students' schools. The schools were instructed to reserve 15 minutes for the students to prepare for the tasks and 15 minutes for recording them (Härmälä et al., 2014). Students' performance was assessed separately for each task by their teachers, who received detailed instructions accompanied by sample performances with model assessments. As a reliability check, 10% of the test performances were assessed also at the Finnish Education Evaluation Centre. The assessment utilised the Finnish application of the CEFR (Finnish National Agency for Education, 2004), which draws on the original CEFR (Council of Europe, 2001). Following the original CEFR scale, it is somewhat vague on fluency but includes the utterance fluency descriptors from the CEFR qualitative aspects scale. The criteria for the proficiency levels relevant for the present study are identical to the CEFR and summarised in Table 1.

Table 1. Utterance fluency criteria considered in the Finnish application of the CEFR

Proficiency level	Mentions of utterance fluency
B2	fairly even tempo, occasional hesitation, few longer pauses
B1	evident pausing and repair
A2	evident pausing, false starts, and reformulation

Note. Adapted from Finnish National Agency for Education (2004).

Audio tracks of the recordings from the national assessment were made available to the author, who selected speech samples from a semi-spontaneous speaking task, in which the students talked about themselves, their hobbies and summer plans to an imaginary exchange student coming to Finland. Approximately 20 seconds were extracted from 60 students, avoiding identifiable content such as names, schools, and hometowns. As such content usually occurred when the students introduced themselves in the beginning of the speaking task, the 20 seconds were extracted from the middle of the task, starting after the introductions. Previously, the speech samples were subjected to global ratings of comprehensibility and accentedness (see Tergujeff, 2021; 2024), which explains the avoidance of identifiable content and short samples.

The 60 Year 9 students who served as participants represented L1 Finnish (n = 30) and L1 Finland-Swedish (n = 30) students of English, aged 15–16. Their L1s were self-reported in a background questionnaire related to the national assessment. In the questionnaire, they indicated that they spoke either Finnish or Finland-Swedish as their only L1 and home language. In addition to L1, participant selection was based on their spoken proficiency in English. Participants across three CEFR levels (A2, B1, B2) were selected: ten participants per proficiency level per L1 group. This enabled investigating the associations between the participants' utterance fluency and their level of spoken proficiency.

Analyses

The speech data were analysed for eleven utterance fluency features: 1. token frequency (i.e., total number of words produced), 2. speech rate (syllables per minute of total time), 3. articulation rate (syllables per minute of speaking time; i.e. with pauses removed), 4. mean length of run (MLR; average number of syllables between silent pauses), 5. silent pauses (number of silent pauses of 0.25 seconds or longer per minute), 6. silent pause length (mean duration of silent pauses in seconds), 7. filled pauses (number of filled pauses per minute), 8. mid-clause pauses (number of mid-clause silent pauses per minute), 9. corrections (number of self-corrected words per minute), 10. repetitions (number of immediately repeated words per minute), and 11. drawls (number of sound elongations of 0.30 seconds or longer per minute). Such measurements have also been used in previous studies, for example in Brown (2006), Ginther et al. (2010), Iwashita et al. (2008), and Kang and Yan (2018). The measurements were made in collaboration between the author and a trained research assistant, utilising technological tools such as Praat (Boersma & Weenink, 2020), an automatic pause detection script (Lennes, 2006), and a tool that counted syllables out of transcripts of the speech samples. Pauses determined by the script were manually checked and adjusted, and the syllable count was also checked. All the measurements will be published as a data set through the Language Bank of Finland by FIN-CLARIN (www.kielipankki.fi). They are also presented in another study that focusses on how utterance fluency and segmental accuracy are associated with L2 comprehensibility and accentedness (Tergujeff-Vasu, forthcoming).

Descriptive statistics were used to determine mean values for each of the eleven utterance fluency measurements per speakers' CEFR level (A2, B1, B2). Then, the data was tested for normality of distribution and homogeneity of variances with Shapiro-Wilk and Levene's test, respectively. Homogeneity was confirmed, but six measurements proved to lack normal distribution: silent pause length, filled pauses, mean length of run, repetitions, corrections, and drawls. Despite the partial lack of normal distribution, a one-way analysis of variance (ANOVA) was conducted to discover whether the means were significantly different across the three proficiency levels. ANOVA does not assume normal distribution throughout the data, but to double-check, a non-parametric Kruskal-Wallis test was also conducted. As the ANOVA led to the same conclusions as the Kruskal-Wallis test, ANOVA was confirmed as a suitable method of analysis. Effect sizes of the possible differences were measured using eta squared ($\eta 2$), which was interpreted as small if >.06, medium if >.16 and large if >.36. These cut-points were adapted from Plonsky and Oswald's (2014) field-specific cut-points for correlation coefficients within L2 research. According to the authors, .25 correlation coefficients should be interpreted as small, .40 medium, and .60 large. Turning these into R2 values (.06, .16, and .36) indicates how much of the variance in the independent variable can be attributed to the dependent variable: 6% for small effect, 16% for medium effect, and 36% for large effect. As eta squared also indicates the proportion of variance in the independent variable attributed to the dependent variable, the same cut-points were adopted. Finally, a post-hoc Tukey comparison was conducted for a more detailed analysis of differences between the three proficiency levels, because speech features may vary significantly across the levels overall without distinguishing adjacent proficiency levels (e.g., Kang & Yan, 2018).

Results

After completing the eleven utterance fluency measurements, descriptive statistics were drawn on to study mean values across proficiency levels. These comparisons yielded varied results. Increasing patterns were found regarding token frequency, mean length of run, and speech rate and articulation rate, meaning that speakers produced more words and longer stretches of speech, and spoke faster than speakers at the adjacent, lower proficiency level. Declining patterns were found regarding the frequency of all pause types and silent pause length: observing the mean values of these features across the three proficiency levels reveals that the frequency of pauses declines, and silent pauses turned shorter with increased proficiency. Repetition, corrections, and drawls were overall rare in the data. While the frequency of corrections seemed to increase with proficiency, repetitions and drawls did not show a linear pattern. Mean values across proficiency levels are presented in Table 2, together with standard deviation, one-way ANOVA, statistical significance, and effect size measured by eta squared.

Table 2. Typical utterance fluency features across proficiency assessments

Utterance fluency feature	**A2 mean**	**SD**	**B1 mean**	**SD**	**B2 mean**	**SD**	***F***	***p***	η^2
token frequency	23.25	6.17	31.60	7.51	35.40	7.46	15.434	<.001	.35
speech rate	95.04	23.37	131.99	37.25	159.10	31.26	21.308	<.001	.43
articulation rate	177.29	36.46	204.24	30.20	217.14	30.30	7.851	<.001	.22
silent pauses	21.98	5.10	21.51	5.01	19.85	4.26	1.090	.343	.04
silent pause length	1.18	0.43	0.87	0.24	0.72	0.23	10.928	<.001	.28
filled pauses	6.07	5.44	6.05	4.91	4.78	5.25	.402	.671	.01
mid-clause pauses	9.71	4.25	9.02	5.98	5.93	4.18	3.409	.040	.11
mean length of run	4.55	1.71	7.31	5.40	8.86	3.33	6.623	.003	.19
repetitions	0.52	1.35	0.80	1.43	0.62	1.55	.197	.822	.01
corrections	1.30	1.90	2.30	2.83	2.50	2.41	1.430	.248	.05
drawls	1.04	2.03	2.86	3.62	2.49	3.33	1.959	.150	.06

Note. Cut-off point for statistical significance is set at .05, and eta squared is interpreted as large when >.36, medium when >.16, and small when >.06.

To explore whether the found patterns are statistically significant, one-way ANOVAs were conducted. In addition, eta squared values were calculated to measure effect sizes, as statistical significance only determines whether the difference can be explained by chance; eta squared reveals the proportion of variance in the proficiency assessment that can be attributed to the utterance fluency measurement. As Table 2 shows, several utterance fluency features differed statistically significantly across proficiency levels: token frequency, speech and articulation rate, silent pause length, mid-clause pauses, and mean length of run can be interpreted to be linked with the speakers' proficiency level. Out of these, mid-clause pauses demonstrated a small effect (η^2 = .11), while mean length of run, articulation rate, silent pause length and token frequency were measured to have medium effects (η^2 = .19–.35). Speech rate yielded a large effect (η^2 = .43). These results mean that 11% of the variance in the proficiency assessments can be attributed to mid-clause pause frequency, 19% to mean length of run, 22% to articulation rate, 28% to silent pause length, 35% to token frequency, and 43% to speech rate.

The post-hoc Tukey test revealed that speech rate was significantly different across the adjacent proficiency levels: between A2 and B1 (p = .001,) and B1 and B2 (p = .021). Then again, token frequency, articulation rate and silent pause length were significantly different only between the two lower proficiency levels A2 and B1 (p = .001 for token frequency, p = .029 for articulation rate, and p = .011 for silent pause length). As for mid-clause pauses, no statistically significant differences were found between adjacent proficiency levels but only between A2 and B2 (p = .045). The same applies to mean length of run: only the difference between A2 and B2 proved statistically significant (p = .002).

Returning to the ANOVA, the analyses show that five utterance fluency features do not differ significantly across proficiency levels, but any of the patterns observed from the mean values can be explained by chance. Based on the present data, no confirmation was found for pause frequency – whether silent or filled pauses – being linked with L2 proficiency assessments. As for the remaining features (repetitions, corrections, and drawls), it seems that speakers with the lowest proficiency used less of these features compared to the more proficient ones, but the development was not linear.

Discussion

The link between L2 proficiency assessments and speed fluency had been demonstrated in several previous studies (e.g., Ginther et al., 2010; Handley & Wang, 2024; Iwashita et al., 2008; Kang & Yan, 2018; Tavakoli et al., 2020). The present study further verified the link: faster speech rate and articulation rate were linked with higher proficiency assessments, even though slow/fast/suitable tempo as such is not mentioned in the assessment criteria. Similarly, token frequency increased; that is, more proficient speakers produced more words in their speech samples than less proficient speakers. This supports previous findings by Kang & Yan (2018) and Brown (2006). In sum, previous research already had convincing evidence regarding speed fluency and token frequency, and the present study does not contradict the evidence. The results suggest that raters consider tempo and the amount of speech produced as constituents of fluency – affecting L2 proficiency assessments – even if these features are not explicitly included in the assessment criteria. Tavakoli et al. (2020) have suggested that human assessments are at least partly subjective because of this; raters reflect what they intuitively consider to be fluent speech.

Previous research had not found much evidence about hesitation and repairs being significantly linked with proficiency assessments. This is also the case with the present study: repetitions, corrections and drawls did not prove to be associated with proficiency assessments. This finding is of interest, as the assessment criteria mention hesitation such as false starts and reformulation: hesitation is characteristic of levels A1 and A2 and continues to be evident at level B1 but only occasional from level B2 (Finnish National Agency for Education, 2004). In light of the assessment criteria, B2-level speakers could have made less corrections, repetitions, and drawls than the other speakers, but that was not the case. On the contrary, the tendency for these features was to increase along with proficiency level (cf. Peltonen, 2018). However, there were very few hesitations like these in the data; perhaps too few to offer a credible analysis of the link between hesitation and proficiency assessments. Moreover, hesitation decreases from B2 toward full fluency in level C proficiency, according to the assessment criteria. To fully grasp the association between hesitation and proficiency assessments, the range of speakers should be extended to level C proficiency, which was not possible with the present group of participants.

As for the other utterance fluency features that yielded mixed results in previous studies, links to proficiency assessments were verified regarding silent pause length and mean length of run (MLR). Silent pause length was statistically different across the proficiency levels with a medium effect, pauses getting shorter along with increased proficiency. This finding supports the results of Ginther et al. (2010) and Tavakoli et al. (2020), who found a similar pattern. The assessment criteria do not mention anything about pause length nor do they make a difference between the occurrence of pausing at levels A2 and B1, whereas at level B2, only a few longer pauses occur, according to the criteria. Consequently, one would expect the difference in pause length to be found between B1 and B2, but it was located between A2 and B1. For MLR, the results are in line with several previous studies (Ginther et al., 2010; Kang and Yan, 2018; Tavakoli et al., 2020). Mean length of run developed across proficiency levels with medium effect; more advanced speakers spoke in longer runs. Hence, it seems that length of runs may be associated with L2 proficiency assessments.

Even though MLR is tightly connected to the occurrence of silent pauses, silent pause frequency was not found associated with proficiency assessments in the present study. This finding is similar to several previous studies (Brown, 2006; Handley & Wang, 2024; Kang and Johnson, 2018; Tavakoli et al., 2020), but contradicts the assessment criteria, as B2-level speakers are described to have only few longer pauses, while the lower levels of proficiency are characterised by evident pausing. Similarly, frequency of filled pauses did not prove statistically significant across proficiency levels. Previously, Kang and Johnson (2018) found a weak correlation between proficiency assessments and filled pauses, but they conclude that filled pauses may not be a strong indicator of proficiency. Based on the present study, attention is drawn more on where the speaker pauses than how often they pause. Namely, the frequency of mid-clause pauses proved significantly different across proficiency levels; however, the effect size was small. This finding is similar to Handley and Wang's (2024). To sum up, pause frequency does not seem to be associated with proficiency assessments in light of the present study, but the study found support for links between proficiency assessments and silent pause length and frequency of mid-clause pauses.

As discussed above, studying the link between proficiency assessments and hesitation phenomena could have yielded more credible results if the participants had not been limited to proficiency levels A2–B2. The same applies to investigating pausing: the assessment criteria describe pauses (along with hesitation) as something characteristic of A1–A2-level speakers, and that frequent pausing continues at B1; only from B2, pausing starts to decrease. However, the assessment criteria can also be interpreted as explanatory to the results: as pausing and hesitation are considered a part of A2–B1-level speech and only start to decrease from level B2, it is natural that no statistically significant differences were found across levels A2, B1 and B2. Yet, focussing on only three proficiency levels is a clear limitation of the present study. As suggested by for example Kang and Yan (2018) and supported by the present results, utterance fluency may fail to distinguish between adjacent proficiency levels, which also speaks for including a wide range of proficiency levels in studies like the present one. This may be possible when utilising high-stakes exams with large numbers of examinees, and I hope to see such research done more in the future. Such studies can also opt for bigger sample sizes, which was not possible in the present study, and this reduced the possibility of generalising the results.

Additional limitations of the present study concern the selection of the speech samples. The proficiency assessments were based on the entire speaking task, whereas the present study utilised a 20-second sample from each speaker. Even though many of the test performances were not much longer than 20 seconds, cutting the performance may have affected the results. With regard to comparing the present investigation to other similar studies, it should be noted that fluency has been defined and operationalised differently in different proficiency scales, which complicates comparisons and the building of a wider understanding of the contribution of utterance fluency to L2 proficiency assessments. Also, utterance fluency may not be considered systematically within the assessment criteria (Harding, 2017; Pennington & Rogerson-Revell, 2019). This is the case with the Finnish application of the CEFR used in the present study (see Table 1), which may have affected the results. In addition, Tavakoli and Skehan (2005) have demonstrated that the speaking tasks themselves can influence the assessments, which further complicates comparisons. Still, the present study can be regarded as a welcome addition to the body of research on the role of utterance fluency in L2 proficiency assessments, strengthening our understanding especially on the importance of how much the speakers being assessed speak, how fast they speak, where they pause and for how long, and how much they speak between the pauses.

Conclusion and implication

This study set out to investigate how typical utterance fluency features are associated with L2 proficiency assessments based on the Finnish application of the CEFR scale (Finnish National Agency for Education, 2004). To investigate this, a research design was developed around proficiency assessments on L1 Finnish and L1 Finland-Swedish teenagers' L2 English. The study verified previous research findings, according to which faster speech and producing more words is associated with higher proficiency assessments. Also, speaking in longer sequences between pauses was linked to higher proficiency assessments, even though previous findings on the issue were mixed. Similarly, pauses turned shorter along with increased proficiency, and pausing in the middle of clauses decreased. Regarding the frequency of silent and filled pauses and hesitation phenomena, no statistically significant differences were found across proficiency levels.

The findings of the present study have implications for L2 assessment, research, and teaching. As the results support earlier findings on various utterance fluency features being associated with L2 proficiency assessments, those features should be consistently considered in language proficiency scales and in automated assessment. Harding's (2017) criticism towards the original CEFR was justified, and after the modifications done to the new version (Council of Europe, 2020), fluency is described in a scale of its own. In the new fluency scale, the description of pausing forms a systematic continuum across the proficiency levels and can therefore be better considered when assessing L2 proficiency. Similar development would be welcome regarding other utterance fluency features as well.

The updated CEFR still refers to tempo in a manner that suggests that an even tempo signals proficient L2 speech. Like several previous studies, the present study supports the conclusion that a *suitable tempo* (measured as speech rate and/or articulation rate) is crucial for fluent speech and connected to L2 proficiency assessments, while there is no convincing evidence in favour of an even tempo contributing to fluency, comprehensibility, or proficiency. Keeping even tempo as an assessment criterion would require solid, research-based justifications. Hence, I encourage future research on the effects of fluctuating tempo on L2 speech fluency, comprehensibility, and proficiency assessments. In any case, the CEFR and other proficiency scales could consider mentioning suitable tempo, because the link between tempo and fluency, comprehensibility, and proficiency assessments has been established quite widely in research studies. Defining a suitable tempo is naturally challenging due to it being context dependent, but I still believe that the concept is viable as an assessment criterion. Associating it with comprehensibility might be one solution: too slow speech could be characterised to cause difficulty in connecting words into thought units and phrases into meaningful entities, whereas speech is too fast if listeners need to struggle to identify key words, for example. Clear and concrete assessment criteria are overall important in L2 assessment, along with efficient and thorough rater training.

As for L2 teaching, the present study implies that speaking fast enough and avoiding long silent pauses especially in the middle of clauses are good goals for learners. These speech features have been found associated with fluency, comprehensibility, and proficiency assessments, and they are hence in line with the Intelligibility Principle, the foundation of modern pronunciation teaching (see Levis, 2005). Learners more often suffer from slow than fast speech, and speaking rate increases along with vocabulary and grammar proficiency. Yet, tempo *per se* can also be practised, and mere drawing of the learners' attention to tempo is a good start. Similarly, raising awareness about the negative effects of long silent pauses on comprehensibility and about where it is natural to pause can be recommended to L2 teaching. I would also welcome pausing-focussed, explicit activities developed for L2 instruction. In general, it is important to view L2 teaching and assessment tightly linked, making sure that both emphasise aspects of language that are crucial for successful communication in the target language.

Acknowledgements

The author is grateful to research assistant Valtteri Nyyssönen for the help with speech data analyses.

Author disclosures

Data utilised in this article was collected for the research project *Intelligibility, Comprehensibility and Accentedness of English Spoken by Finns*, funded by the Research Council of Finland (grant number 315980) and the University of Jyväskylä 2018–2021.

ORCID iD

Elina Vasu https://orcid.org/0000-0002-3867-9915

References

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer [computer program]. www.praat.org

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (eds.), *IELTS research reports 2006* (pp. 71–89). IELTS Australia and British Council. https://ielts.org/researchers/our-research/research-reports/an-examination-of-the-rating-process-in-the-revised-ielts-speaking-test

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European Framework of reference for languages: Learning, teaching, assessment*. *Companion volume*. Council of Europe. http://www.coe.int/lang-cefr

De Jong, N. H., Groenhout, R., Schoonen, R, & Hulstijn, J. H. (2015). Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223–243. https://doi.org/10.1017/s0142716413000210

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://doi.org/10.1075/lllt.42

Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Routledge. https://doi.org/10.4324/9781315165103-1

Finnish National Agency for Education. (2004). *National core curriculum for basic education.* Finnish National Agency for Education.

Finnish National Agency for Education. (2014). *National core curriculum for basic education.* Finnish National Agency for Education.

Gass, S. M., & Selinker, L. (Eds.) (1992). *Language transfer in language learning.* Revised edition. John Benjamins. https://doi.org/10.1075/lald.5?locatt=mode:legacy

Ghanem, R., & Kang, O. (2018). Pronunciation features in rating criteria. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation.* Routledge. https://doi.org/10.4324/9781315170756-7

Ginther, A., Dimova, S. & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379–399. https://doi.org/10.1177/0265532210364407

Handley, Z. L. & Wang, H. (2024). What do the measures of utterance fluency employed in Automatic Speech Evaluation (ASE) tell us about oral proficiency? *Language Assessment Quarterly, 21*(1), 3–32. https://doi.org/10.1080/15434303.2023.2283839

Harding, L. (2017). What do raters need in a pronunciation scale? The users' view. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 12–34). Multilingual Matters. https://doi.org/10.21832/9781783096855-004

Huhta, A., Kallio, H., Ohranen, S., & Ullakonoja, R. (2020). Fluency in language assessment. In P. Lintunen, M. Mutta, & P. Peltonen (Eds.), *Fluency in L2 learning and use* (pp. 129–145). Multilingual Matters. https://doi.org/10.21832/9781788926317-011

Härmälä, M., Huhtanen, M., & Puukko, M. (2014). *Englannin kielen A-oppimäärän oppimistulokset perusopetuksen päättövaiheessa 2013*. Finnish National Agency for Education.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*(2), 301–315. https://doi.org/10.1016/j.system.2010.01.005

Kang, O. & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly, 15*(2), 150–168. https://doi.org/10.1080/15434303.2018.1451531

Kang, O. & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment, 1*, 24–39. https://doi.org/10.23977/langta.2018.11003

Kang, O., Ahn, H., Yaw, K. & Chung, S-Y. (2021). Investigation of relationships between learner background, linguistic progression, and score gain on IELTS. *IELTS Research Reports Online Series, No. 1.* British Council, Cambridge Assessment English and IDP: IELTS Australia. https://ielts.org/researchers/our-research/research-reports/investigation-of-relationships-between-learner-background-linguistic-progression-and-score-gain-on-ielts

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Lehtonen, J. (1979). Speech rate and pauses in the English of Finns, Swedish-speaking Finns, and Swedes. *AFTIL, 6,* 35–51.

Lennes, M. (2006). Praat script mark_pauses.praat. Retrieved Oct 1, 2020, from https://lennes.github.io/spect/.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. https://doi.org/10.2307/3588485

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*(2), 159–182. https://doi.org/10.1111/1467-9922.00038

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*(4), 451–468. https://doi.org/10.1017/S0272263101004016

Paananen-Porkka, M. (2007). *Speech rhythm in an interlanguage perspective: Finnish adolescents speaking English*. University of Helsinki.

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *Modern Language Journal, 102*(4), 676–692. https://doi.org/10.1111/modl.12516

Pennington, M. C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research: Contemporary perspectives*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-47677-7

Pihko, M-K. (1997). *"His English sounded strange." The Intelligibility of native and non-native English pronunciation to Finnish learners of English*. University of Jyväskylä, Centre for Applied Language Studies.

Pinget, A-N., Rutger Bosker, H., Quené, H. & de Jong, N.H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31*(3), 349–365. https://doi.org/10.1177/0265532214526177

Plonsky, L. & Oswald, F.L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Segalowitz, N. (2010). *The cognitive bases of second language fluency.* Routledge. https://doi.org/10.4324/9780203851357

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4). 510–532. https://doi.org/10.1093/applin/amp047.

Tavakoli, P., Nakatsuhara, F., & Hunter, A-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Tergujeff, E. (2021). Second language comprehensibility and accentedness across oral proficiency levels: A comparison of two L1s. *System 100*, 1–11. https://doi.org/10.1016/j.system.2021.102567

Tergujeff, E. (2024). Second language learners listening to their peers' English: Is there a shared L1 language effect for L2 comprehensibility and accentedness? *European Journal of Applied Linguistics, 12*(2), 469–491. https://doi.org/10.1515/eujal-2022-0014

Tergujeff-Vasu, E. (forthcoming). What constitutes second language comprehensibility and accentedness? Evidence from new speaker groups. *Journal of Monolingual and Bilingual Speech, 6*(3).

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1–30. https://doi.org/10.1017/S0272263106060013

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language & Cognition, 15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

Volín, J., & Skarnitzl, R. (2010). The strength of foreign accent in Czech English under adverse listening conditions. *Speech Communication, 52*(11/12), 1010–1021. https://doi.org/10.1016/j.specom.2010.06.009

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Zhang, M., Bridgeman, B., & Davis, L. (2020). Validity considerations for using automated scoring in speaking assessment. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 21–31). Routledge. https://doi.org/10.4324/9781315165103-2